DOCUMENT RESUME

ED 189 191 TM 800 373

AUTHOR Byars, Alvin Gregg
TITLE Practical Procedures for Constructing Mastery Tests to Minimize Errors of Classification and to Maximize or Optimize Decision Reliability.
INSTITUTION Charleston County School District, North Charleston, S.C.
PUB DATE Apr 80
NOTE 41p.: Paper presented at the Annual Meeting of the National Council on Measurement in Education (Boston, MA, April 8-10, 1980). Prepared by the Office of Evaluation and Research. Best copy available.

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Cutting Scores: Difficulty Level: Grade 4: Intermediate Grades: Item Banks: *Mastery Tests: Predictor Variables: *Test Construction: Test Items: Test Reliability: *Test Validity
IDENTIFIERS Beta Binomial Model: *Test Length

ABSTRACT

The objectives of this investigation are to develop, describe, assess, and demonstrate procedures for constructing mastery tests to minimize errors of classification and to maximize decision reliability. The guidelines are based on conditions where item exchangeability is a reasonable assumption and the test constructor can control the number of items on the test and its cutoff score. Eight steps for minimizing errors of classification are described. The major difficulty of the procedure is the identification of an appropriate criterion variable. Suggested is the identification of a reasonable concurrent criterion variable in order to set temporary standards of mastery until sufficient longitudinal data can be obtained. The procedure for maximizing decision reliability involves five steps. A shortcut to the procedure is provided in the form of tables for determining the proportion correct for mastery for tests ranging in length from five to ten items. A major advantage of the procedures described for generating mastery test forms from an item bank is that they are well within the reach of most local school districts desiring accuracy in mastery assessment programs. The procedures do not require unduly complicated computational procedures. (Author/GSK)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED189191

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

BEST COPY AVAILABLE

OFFICE of EVALUATION & RESEARCH

PRACTICAL PROCEDURES FOR CONSTRUCTING MASTERY TESTS TO MINIMIZE ERRORS OF CLASSIFICATION AND TO MAXIMIZE OR OPTIMIZE DECISION RELIABILITY

Alvin Gregg Byars

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
A. Byars

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Paper presented at the NCME annual convention, Boston, 1980.

Lawrence G. Derthick, Jr.
Superintendent

Jerry McGee
Associate Superintendent
Instructional Services

Bill Quesenbery
Director
Eval. & Research

TM 800373

PRACTICAL PROCEDURES FOR CONSTRUCTING MASTERY TESTS TO MINIMIZE ERRORS OF CLASSIFICATION AND TO MAXIMIZE OR OPTIMIZE DECISION RELIABILITY

Alvin Gregg Byars
Evaluation and Research
Charleston County (SC) School District

ABSTRACT

In many objectives-based instructional programs, decisions concerning student mastery or nonmastery of an instructional objective are based on scores obtained from a test referenced to a domain of items specified by the objective. When such tests are constructed by randomly selecting dichotomously scored items from the domain, item exchangeability may be a reasonable assumption. Provided that conditions of exchangeability are met, practical procedures are demonstrated which provide the constructor of mastery tests the ability to 1) select a cutoff score in such a way as to maximize the criterion-related validity and to minimize the errors of classification for a given test length, and then 2) adjust the length of the test, within administrative constraints, to maximize or optimize the decision reliability for the test and to further decrease the errors of classification.

Paper presented at the NCME annual convention, Boston, 1980.

PRACTICAL PROCEDURES FOR CONSTRUCTING MASTERY TESTS TO MINIMIZE ERRORS OF CLASSIFICATION AND TO MAXIMIZE OR OPTIMIZE DECISION RELIABILITY

Alvin Gregg Byars
Evaluation and Research
Charleston County (SC) School District

OBJECTIVES OF INQUIRY

In many objectives-based instructional programs, decisions concerning student mastery or nonmastery of an instructional objective are based on scores obtained from a test referenced to a domain of items specified by the objective. When such tests are constructed by randomly selecting dichotomously scored items from the domain (or from an item bank representing the domain), item exchangeability may be a reasonable assumption (Huynh, 1976b). Under conditions of exchangeability, the accuracy of mastery/nonmastery decisions may be shown to depend on four parameters of the test and test scores: 1) the mean item difficulty, 2) the variability of the test scores, 3) the length of the test, and 4) the cutoff score or proportion of items which must be answered correctly to obtain mastery status (Huynh, 1976a, 1976b, 1977).

The mean item difficulty and the score variability are determined by the characteristics of the population to which students belong about whom decisions are to be made (Thorndike, 1971). Consequently, the number of items on the test and the cutoff score for the test appear to be the parameters over which the test constructor has the most control in attempts to improve the accuracy of instructional decisions.

Accurate classification decisions must be both valid, with validity assessed by some criteria external to the test (Garcia-Quintana, 1979; Huynh, 1979; Huynh and Saunders, 1979), and reliably consistent, with consistency assessed by a variety of methods (Subkoviak, 1978).

The objectives of the investigation were to develop, describe, assess and demonstrate practical procedures for constructing mastery tests to minimize errors of classification and to maximize or optimize decision reliability, under conditions where item exchangeability is a reasonable assumption and where the test constructor can control the number of items on the test and the cutoff score for the test.

PROCEDURES FOR DETERMINING THE PROPORTION CORRECT FOR MASTERY

Minimizing Errors of Classification

Step 1:

Construct a sample form of the mastery test by randomly selecting items from the bank of items for the corresponding objective. Approximately 10 items is a practical number with which to work and should provide reasonably accurate estimates.

Step 2:

Identify an appropriate criterion for judging the validity of the test.

If student mastery of an instructional objective is of interest in itself, without regard for subsequent instructional objectives which may or may not be related, then concurrent validity would appear to be the appropriate form of criterion-related validity. Examples of concurrent criterion measures include teacher judgement of student mastery/nonmastery; and success or failure on other previously validated mastery tests.

If student mastery of an instructional objective is of interest within some larger context, and if mastery of the objective has implications for mastery of some subsequent task, then predictive validity would appear to be the appropriate form of criterion-related validity. Examples of predictive criterion measures include performance in subsequent and presumably more complex instructional units; performance in subsequent, more advanced, instructional programs; and performance on a subsequent

examination for competency certification.

Step 3:

Administer the sample test to an appropriate sample of students and obtain test scores for each student. What constitutes an appropriate sample of students must be determined, of course, within the context of the overall test development strategy.

Step 4:

Obtain estimates of mastery/nonmastery on the criterion variable for each student in the sample. If desired, obtain an estimate of certainty for each mastery/nonmastery estimate. For example, if teacher judgement is used as the criterion, an estimate of the degree of certainty for each judgement may be obtained. If performance on a subsequent competency examination is used as the criterion, a probability of mastery might be obtained for each score on the exam. To avoid contamination, the criterion measures should be obtained independently of the mastery test scores.

Step 5:

As illustrated in FIGURE 1, operationally define a student's "true" mastery status as his status on the criterion variable. Operationally define a student's "observed" status as master if he scores at or above a given cutoff score, or as nonmaster if he scores below the given cutoff score.

Step 6:

As formulated in TABLE 1, compute the proportions of true positive, true negative, false positive and false negative decisions for each possible cutoff score on the test. Note the different formulations for the weighted and unweighted methods.

Step 7:

For each possible cutoff score on the test, compute the P index of decision consistency, or proportion of "correct" decisions. The P index is computed in the same way for the weighted and unweighted methods as formulated in TABLE 1.

Step 8:

Determine the proportion correct for mastery which maximizes the criterion-related validity of the test, and minimizes the errors of classification, by selecting the cutoff score for which the P index is largest.

It is generally recognized that, for instruction which is highly structured and hierarchical in nature, a false positive decision may be more detrimental to the student than a false negative decision. Although a false negative decision may slow the student's progress through the instructional process, a false positive decision may deny the student the opportunity to receive instruction on unmastered objectives. If the test constructor believes this to be the case, an additional constraint may be imposed in determining the proportion correct for mastery. A cutoff score would be selected such that the P index is maximum,

given that the proportion of false positive decisions is less than the proportion of false negative decisions. In such a case, the test constructior may be justified in sacrificing some criterion-related validity in order to decrease the proportion of false positive decisions.

FIGURE 1

OPERATIONAL DEFINITIONS OF

"TRUE" AND "OBSERVED" MASTERY STATUS

"OBSERVED" STATUS: Status on Mastery Test	"TRUE" STATUS: Status on Criterion Variable: Nonmaster	"TRUE" STATUS: Status on Criterion Variable: Master
Master	P_{10}	P_{11}
Nonmaster	P_{00}	P_{01}

"True" mastery status is operationally defined as the status on the criterion variable.

"Observed" mastery status is operationally defined as the status on the mastery test.

P_{11} = proportion of true positive decisions.

P_{00} = proportion of true negative decisions.

P_{10} = proportion of false positive decisions.

P_{01} = proportion of false negative decisions.

TABLE 1

SUMMARY OF COMPUTATIONS FOR ESTIMATING

CRITERION-RELATED VALIDITY

DESCRIPTION	COMPUTATIONAL FORMULA
Compute the unweighted proportion of true positive decisions, where N is the number of students in the sample.	$P_{11} = \sum_{i=1}^{N} \Psi_i / N$ $\Psi_i =$ 1 if master on test and criterion variable, 0 otherwise
Compute the unweighted proportion of true negative decisions.	$P_{00} = \sum_{i=1}^{N} \Psi_i / N$ $\Psi_i =$ 1 if nonmaster on test and criterion variable, 0 otherwise
Compute the unweighted proportion of false positive decisions.	$P_{10} = \sum_{i=1}^{N} \Psi_i / N$ $\Psi_i =$ 1 if master on test and nonmaster on criterion variable, 0 otherwise
Compute the unweighted proportion of false negative decisions.	$P_{01} = \sum_{i=1}^{N} \Psi_i / N$ $\Psi_i =$ 1 if nonmaster on test and master on criterion variable, 0 otherwise

TABLE 1 (continued)

DESCRIPTION	COMPUTATIONAL FORMULA
Compute the weighted proportion of true positive decisions, where ω_i is the weight for the ith student on the criterion variable.	$P_{11} = \frac{\sum_{i=1}^{N} \omega_i \Psi_i}{\sum_{i=1}^{N} \omega_i}$ Ψ_i is defined as for the unweighted method
Compute the weighted proportion of true negative decisions.	$P_{00} = \frac{\sum_{i=1}^{N} \omega_i \Psi_i}{\sum_{i=1}^{N} \omega_i}$ Ψ_i is defined as for the unweighted method
Compute the weighted proportion of false positive decisions.	$P_{10} = \frac{\sum_{i=1}^{N} \omega_i \Psi_i}{\sum_{i=1}^{N} \omega_i}$ Ψ_i is defined as for the unweighted method.
Compute the weighted proportion of false negative decisions.	$P_{01} = \frac{\sum_{i=1}^{N} \omega_i \Psi_i}{\sum_{i=1}^{N} \omega_i}$ Ψ_i is defined as for the unweighted method.

TABLE 1 (continued)

DESCRIPTION	COMPUTATIONAL FORMULA
Compute the P index as an estimate of the criterion-related validity for the test.	$P = P_{11}+P_{00}$

DELIMITATIONS AND DISCUSSION

Larger samples of items from the bank would be expected to provide more accurate estimates. Ten items is simply a convenient number. A reasonable estimate of the reliability of the obtained test score distribution might be computed in the form of a split-half internal consistency estimate. Other score reliability estimates might also serve this purpose (see Thorndike, 1971).

The major difficulty of the procedure is the identification of an appropriate criterion variable. If instruction is to be relevant to success in later life, then, certainly, the appropriate form of criterion-related validity would be predictive validity. However, longitudinal data for establishing predictive validity is often unavailable. A compromise strategy for the test constructor might be the identification of a reasonable concurrent criterion variable in order to set temporary standards of mastery until sufficient longitudinal data can be obtained.

When measurement of the criterion variable involves judgement, estimates of mastery/nonmastery are, at best, crude. When weighting is used to refine these estimates, the weights themselves may be rather crude, particularly when the weights are based on the ordered degree of certainty of each judgement. However, the weighting strategy should provide better estimates than the strategy of arbitrarily assigning unit weighting to each judgement as is done in the unweighted method.

An alternative, and perhaps preferred, strategy might be to obtain estimates of mastery from two or more judges for each student. This would allow the computation of a reliability, or inter-rater agreement, index for the criterion measures. However, this is often unfeasible in the typical

instructional setting.

Another alternative is to classify students into a trichotomy, master, uncertain and nonmaster, on both the test and the criterion (Garcia-Quintana, 1979; Huynh, 1979). This method provides a range of cutoff scores and allows the test constructor some flexibility in controlling the relative proportions of false positive and false negative decisions. However, this method typically deflates the P index considerably. This is simply because a two-way dichotomy requires the data to be fit into four distinct categories, two of which are used in computing P; a two-way trichotomy requires the data to be fit into nine distinct categories, only three of which are used in computing P.

When teachers are both administering the mastery test and providing judgements of student mastery, it is likely that the test scores and criterion ratings will not be independent. This problem can be reduced to some extent by obtaining the judgements prior to administering the test, and by following strict procedures for test administration.

When the distribution of ability is not fairly continuous in the student sample, as often occurs when the sampling unit is rather large, two or more cutoff scores may have equally high P indices and equal proportions for each of the four decision categories. In such a case, the "best" cutoff score is not readily apparent. For example, if the sampling unit is a classroom, this problem usually occurs when the within-class test variances are much smaller than the across-class test variance.

This is usually much less of a problem when the weighted method is used. This is simply because we are introducing more variance into the criterion measures. P indices obtained by the weighted method typically tend to be higher than those obtained by the unweighted method.

The estimate of the proportion correct for mastery which maximizes criterion-related validity should, of course, be a reliable or consistent estimate. Several methods might be employed to assess the reliability of the estimates. One reasonable method would be to apply the procedure to two or more subsamples of students. The P indices for each possible cutoff score could then be ranked for each subsample. A coefficient of concordance, W, could be computed for the rankings to provide an estimate of consistency or inter-subsample reliability (Siegel, 1956).

PROCEDURES FOR DETERMINING THE "BEST" TEST LENGTH

Maximizing/Optimizing Decision Reliability

Step 1:

Applying the formulas given in TABLE 2, compute the alpha, α, and beta, β, parameters of the beta-binomial model for the sample mastery test form.

Step 2:

For each possible test length, ranging from two items to the total number of items in the bank, compute the univariate probability density function (pdf) for the beta-binomial model; and compute the bivariate beta-binomial pdf. These computations are formulated in TABLE 2.

Step 3:

For each possible test length, select a cutoff score corresponding to the previously determined proportion correct for mastery. If the proportion correct for mastery does not correspond to an integer cutoff score, one of two rounding methods may be used. If the relative proportions of false positive and false negative decisions are of little concern, simply round the cutoff score to the nearest integer. If the test constructor believes that the false positive decision is the more detrimental of the two false decisions, the cutoff score should be rounded to the next higher integer. With this latter procedure, the test results will tend to err more toward false negative decisions.

Step 4:

For each possible test length and corresponding cutoff score, compute the P index for decision consistency as formulated in TABLE 2. This P index may be interpreted as the proportion of consistent mastery decisions expected across multiple test administrations.

Step 5:

In order to determine the "best" test length for a given purpose, select the shortest test length which will provide adequate reliability for the decision to be made. For example, if group decisions are to be made, a decision reliability of .60 or .70 may be considered adequate and only a very few items may be required. If individual decisions are to be made, a decision reliability of .85 or .90 might be preferred. Consequently, more items would be required. This procedure allows the selection of an adequate number of items without selecting more items than are necessary.

TABLE 2

SUMMARY OF CALCULATIONS FOR THE BETA-BINOMIAL MODEL

DESCRIPTION	COMPUTATIONAL FORMULA
Compute the mean for the sample test score distribution, where N is the number of students in the sample and X_i is the score for the ith student.	$\bar{x} = \sum_{i=1}^{N} X_i/N$
Compute the standard deviation for the sample test score distribution.	$s = \sqrt{\frac{\sum_{i=1}^{N} (X_i-\bar{x})^2}{N-1}}$
Compute the KR21 reliability estimate where n is the number of items on the test.	$\alpha_{21} = \frac{n}{n-1}(1- \frac{\bar{x}(n-\bar{x})}{ns^2})$
Compute the alpha and beta parameters of the beta-binomial model.	$\alpha = (-1+1/\alpha_{21})\bar{x}$ $\beta = -\alpha+n/\alpha_{21}-n$
Compute the univariate beta-binomial score density for each score, x.	For small test scores: $f(0) = \prod_{i=1}^{n} \frac{n+\beta+i}{n+\alpha+\beta-i}$ $f(x+1) = f(x) \cdot \frac{(n-x)(\alpha+x)}{(x+1)(n+\beta-x-1)}$, $x=0,1,\ldots,n-1$ For large test scores: $f(n) = \prod_{i=1}^{n} \frac{n+\alpha-i}{n+\alpha+\beta-i}$

TABLE 2 (continued)

DESCRIPTION	COMPUTATIONAL FORMULA
	$f(x-1) = f(x) \cdot \frac{x(n+\beta-x)}{(n-x+1)(\alpha+x-1)}$, $x=1,\ldots,n$
Compute the bivariate beta-binomial density for each score.	For small values of x and y: $f(0,0) = \prod_{i=1}^{2n} \frac{2n+\beta-i}{2n+\alpha+\beta-i} = f(0) \cdot \prod_{i=1}^{n} \frac{2n+\beta-i}{2n+\alpha+\beta-i}$
	$f(x+1,y) = f(x,y) \cdot \frac{(n-x)(\alpha+x+y)}{(x+1)(2n+\beta-x-y-1)}$
	For large values of x and y: $f(n,n) = \prod_{i=1}^{2n} \frac{2n+\alpha-i}{2n+\alpha+\beta-i} = f(n) \cdot \prod_{i=1}^{n} \frac{2n+\alpha-i}{2n+\alpha+\beta-i}$
	$f(x-1,y) = f(x,y) \cdot \frac{x(2n+\beta-x-y)}{(n-x+1)(\alpha+x+y-1)}$
Compute P, the raw agreement index for multiple administrations, for the given cutoff score, c.	For cutoff scores near 0: $P_0 = \sum_{x=0}^{c-1} f(x)$
	$P_{00} = \sum_{x,y=0}^{c-1} f(x,y)$

TABLE 2 (continued)

DESCRIPTION	COMPUTATIONAL FORMULA
	$P = 1-2(P_0-P_{00})$
	For cutoff scores near n:
	$P_1 = \sum_{x=c}^{n} f(x)$
	$P_{11} = \sum_{x,y=c}^{n} f(x,y)$
	$P = 1-2(P_1-P_{11})$

20

Adapted from Byars (1979).

DELIMITATIONS AND DISCUSSION

The beta-binomial model requires the relatively mild assumption of item exchangeability, that is, the items should be content homogeneous and approximately equal in difficulty. Empirical studies (Huynh, 1978a Subkoviak, 1978) have demonstrated that the model is quite robust to gross violations of the assumption of equal item difficulties. This assumption appears very reasonable for tests of mastery of a specific instructional objective when items are randomly drawn from a bank. The beta-binomial model is generally suited to a wide variety of empirical distributions, including U-shaped, inverted U-shaped, J-shaped and reverse J-shaped (Huynh, 1978b). However, caution should be exercised in applying the model to the data without assessing the "fit" of the model. This is particularly true for distributions which are not unimodal.

A simple chi-square test of fit (Siegel, 1956) can be performed to assess the fit of the observed sample test distribution to the predicted distribution derived from the beta-binomial univariate pdf. Unless the distribution of ability is highly heterogeneous and the sample size is fairly large, this method may require collapsing the score categories on one or both ends of the distribution to obtain sufficiently high expected frequencies. Consequently, chi-square may provide a rather crude test of fit.

Of course, the beta-binomial is not the only model that could be used for this purpose. For example, see Subkoviak (1978). However, the beta-binomial model does have the advantages of requiring only a single test administration and of providing estimators which are fairly consistent with those obtained for multiple administrations (Huynh, 1979 ; Subkoviak, 1978).

Beta-binomial estimates of P do tend to be somewhat conservative when compared with estimates obtained by other methods. In general, the shorter the test, the more conservative the estimate of P will be.

A shortcut to the procedure described here has been provided by Huynh (1978b) in the form of tables for determining P for tests ranging in length from five to 10 items. These tables also provide estimates of the standard error of estimate for P.

A major advantage of the procedures described for generating mastery test forms from an item bank is that they are well within the reach of most local school districts. All computations for this study were performed on a relatively small microcomputer (IBM 5100, 32K). This researcher's experience has shown that maintenance of a large item bank and generation of mastery tests can easily be handled by inexpensive and quite small (16K or more) computing systems. Such a system can provide the local school district with a high degree of flexibility and control of accuracy in mastery assessment programs.

EDUCATIONAL IMPORTANCE OF THE STUDY

The implications of the described procedures for improving the accuracy of educational measurement and instructional decision making should be readily apparent. The procedures are practical, easily comprehended, and do not require unduly complicated computational procedures. Furthermore, it has been demonstrated that the procedures are well within the reach of local school districts desiring some control of flexibility and accuracy in mastery assessment programs. Provided that a few simple and mild assumptions are met, the procedures provide the constructor of mastery tests the ability to 1) select a cutoff score in such a way as to maximize the criterion-related validity and to minimize the errors of classification for a given test length, and then 2) adjust the length of the test, within administrative constraints, to maximize or optimize the decision reliability for the test and to further decrease the errors of classification.

EXAMPLE

An example of the application to these procedures follows. The instrument used in the example was extracted from the first completed segment of a large scale cognitive assessment system in the Language Arts which is currently under development by the Charleston County (SC) School District. The test was designed to measure skill in recognizing affixes and was one of several tests assessing elementary word recognition skills.

Test Development

Procedures for constructing the tests were designed to assure that test score interpretations possessed various forms of validity. Test development will be outlined in terms of the validity issues addressed.

To a major extent, face validity and logical content validity were built into each test during the item construction and item/objective revision phases of development. Item writers, selected for their content expertise, developed domain specifications from each instructional objective which contained:

* a statement of the objective in behavioral content terms,
* a set of teacher and/or student directions for the test,
* a sample test item, and
* the correct response and examples of distractors for the sample item.

See FIGURE 2 for the domain specifications for the example test.

Items were constructed from the domain specifications until the domain was exhausted or until an adequately representative sample of the domain had been obtained. Twenty items were written for the example test.

Empirical content validity was assessed by test-objective congruence procedures (Anderson, 1977). The proportion of judges, on a twelve member panel of content specialists, matching a test with its intended objective was computed as an index of empirical content validity for the test. The empirical index of content validity for the example test was 1.00.

To some extent, rational construct validity was assured during the procedures for assuring and assessing content validity. Construct validity was empirically assessed through application of the Rasch model (Wright, 1977), or one-parameter latent trait model, to data obtained during pilot testing.

The proportion of items on a test which "fit" the model was computed as an index of empirical construct validity. This index represents the extent to which a student's score on the test (with items dichotomously scored) may be directly interpreted in terms of the student's ability with respect to the instructional objective. For the example test, the empirical index of construct validity was 1.00.

Rasch ability and difficulty estimates were used in an overall scaling, linking and data simulation procedure for the entire assessment system.

For pilot testing, a form of the test was constructed by randomly selecting 10 items from the test bank of 20 items. See FIGURE 3.

Obtaining Pilot Data

The procedures used to select student samples for pilot testing were developed primarily for efficiency in scaling and linking within the overall assessment system. Consequently, the sample of students receiving any individual test was not necessarily representative of the student population. Since these student selection procedures are not central to the theme of this paper,

they will not be discussed here. Subsequent to the linking and scaling procedures, 1000 cases were simulated for each grade for each test to improve the representativeness of the computed statistics for the intended student population. The data presented here were obtained from 89 fourth grade students from three classrooms receiving the example test.

Teacher judgement was selected as the criterion variable for determining the proportion correct for mastery. Student performance on some subsequent instructional objective for which the illustrated objective was prerequisite would have been a preferred criterion measure. However, teacher judgement was judged to be an adequate concurrent criterion for setting temporary standards until sufficient longitudinal data could be obtained to identify adequate predictive criterion measures.

Teachers were asked to classify each student as "master" or "nonmaster" on the corresponding instructional objective and to indicate whether they were "uncertain", "fairly certain" or "very certain" of each classification. See FIGURE 4. Teacher judgements were obtained prior to the administration of the test in order to reduce the problem of dependence between the judgements and the test scores.

Results

Several statistics were computed for each test from the pilot data. However, many of these statistics were used in the construction of survey tests for purposes of norm-referenced interpretations and are not relevant to the theme of this paper. Consequently, these statistics will not be discussed here.

Both the unweighted and weighted procedures for determining the proportion correct for mastery were carried out. For the weighted procedure, teacher judgements indicated to be uncertain, fairly certain and very certain were assigned weights of one, two and three, respectively. Results of the unweighted procedure are presented in TABLE 3. Results of the weighted procedure are presented in TABLE 4.

The Kuder-Richardson formula 21 estimate of internal consistency was computed as an estimate of the reliability of the obtained sample test score distribution. For the example test, this estimate was .5523.

The data presented in TABLES 3 and 4 are based on the combined data for the three classes. In order to obtain some indication of the inter-class reliability of the obtained P indices for the various cutoff scores, the unweighted and weighted procedures were carried out separately for each of the three classes. The P indices for each cutoff score were ranked for each class. Kendall's coefficient of concordance, W, was then computed for the ranks. For the unweighted procedure, W was equal to .5068 (probability $< .20$); for the weighted procedure, W equalled .7754 (probability $< .02$).

From an examination of TABLES 3 and 4, it may be seen that both procedures indicate a proportion correct for mastery of .90 for maximizing the criterion-related validity for the test. However, the weighted procedure suggests a slightly higher degree of validity; .7921 versus .7640.

This researcher has noted at least three general tendencies in applying the two procedures:

1) the rank ordering of the P indices for the two procedures tends to be the same;

2) the range of P for the weighted procedure tends to be greater than for the unweighted procedure; and

3) the weighted procedure tends to discriminate between various cutoff scores better than does the unweighted procedure.

The first two observations are clearly illustrated in TABLES 3 and 4. Note that these are simply observations and not statements of statistical fact.

It is interesting to note that, for the example test, four different cutoff scores might be considered to fall within an acceptable range of validity (approximately .60 to .80); these cutoff scores are .70, .80, .90 and 1.00. Assuming that a validity index of .60 is adequate, the test constructor could have a considerable degree of flexibility in choosing a cutoff score to control the relative proportions of false positive and false negative decisions. A cutoff of .70 would be expected to yield a loss ratio (proportion of false positive decisions divided by proportion of false negative decisions) of approximately 4.0, while a cutoff of 1.00 would be expected to yield a loss ratio of 0.0.

Prior to applying the procedures for determining the "best" test length, the beta-binomial alpha and beta parameters were computed for the observed test score distribution and a chi-square test of fit was performed between the observed distribution and the distribution predicted by the beta-binomial model. This required collapsing scores on the lower end of the distribution to obtain adequate expected frequencies in all cells. The obtained chi-square was 6.8702 (d.f. = 9, probability $>$ .50). Consequently, the fit was judged to be adequate.

Data obtained from application of the procedure to the sample data is presented in TABLE 5. Assuming that a decision reliability of approximately .70 is adequate for group decisions and that a reliability of .85 is adequate for individual decisions, it may be seen that only four items would be necessary

for a group test and that seven or eight items are probably sufficient for individual tests. Note that cutoff scores in TABLE 5 were rounded up to decrease the proportion of false positive decisions.

In TABLE 6, data are presented which were obtained for 1000 simulated cases. Based on this data, it appears that four or five items would be sufficient for both group and individual decisions.

FIGURE 2

DOMAIN SPECIFICATION FOR WS6.0B

OBJECTIVE: WS6.0B - Inflectional Endings (Affixes):

Given an incomplete sentence, the learner will identify the correct affixed word to complete the sentence.

DIRECTIONS: Read the incomplete sentence in each box. Darken the circle under the word that correctly completes the sentence. Let's do the sample.

SAMPLE: The _____ of the furniture made the room look larger.

prearrangement	subarranging	rearrangement
(a)	(b)	(c)

(The correct answer is "rearrangement".)

FIGURE 3

Structural Analysis

Target Objective Code WS 6.0b

Process Objective Code _____

P _____

I X

P-I _____

Directions:

Read the incomplete sentence in each box. Darken the circle under the word that correctly completes the sentence. Let's do the sample.

SAMPLE: The _____ of the furniture made the room look larger.

prearrangement (a) subarranging (b) rearrangement (c)

Now do the test.

1. The _____ of Mr. Jones seemed strange to everyone.

reappeared (a) disappearance (b) misappearment (c)

2. We want an _____ of the baby's picture.

enlargeness (a) imlargement (b) enlargement (c)

3. The road was _____ because of the car accident.

nonpassive (a) unpassion (b) impassable (c)

FIGURE 3 (continued)

4. It takes the _____ of all the members to make a team.

cooperation	disoperation	inoperable
(a)	(b)	(c)

5. Jim spoke _____ to his new neighbor.

impolitely	nonpolitely	mispolitely
(a)	(b)	(c)

6. It is _____ to come to school without our homework.

inexcusable	unexcusable	imexcusable
(a)	(b)	(c)

7. The _____ of the football player's shoulder caused a delay in the game.

mislocation	unlocatable	dislocation
(a)	(b)	(c)

8. It was _____ to have a picnic during the thunderstorm.

impossible	inpossible	unpossible
(a)	(b)	(c)

9. We had a _____ snack by the pool.

nonday	midday	preday
(a)	(b)	(c)

10. The boys made a _____ around the fire and told stories.

midcircle	bicircle	semicircle
(a)	(b)	(c)

FIGURE 4

DIRECTIONS FOR COMPLETING THE STUDENT MASTERY RATING FORM

1) Write the names of the students in your class on the attached rating form in alphabetical order. Be certain that the names appear in the SAME ORDER ON ALL RATING FORMS. The number beside each student's name will be used to identify the student on each of the student's answer sheets.

2) Make certain that the identification code for the objective below is the same as the code which appears next to "SUBJECT" at the top of the attached rating form.

3) Carefully read the objective below:

OBJECTIVE WS 6.0

Inflectional Endings: (Affixes)	Given an incomplete sentence, the learner will identify the correct affixed word to complete the sentence.

4) Next to each name on the rating form, circle "a" for MASTER if you believe that the student has mastered the objective. Circle "b" for NONMASTER if you believe that the student has not mastered the objective.

5) After circling "a" or "b" to rate the student on the objective, then circle "c", "d" or "e" to indicate how certain you are of your rating. You should circle "c" if you are UNCERTAIN, circle "d" if you are FAIRLY CERTAIN, and circle "e" if you are VERY CERTAIN of your rating.

6) When you have completed the above, darken the bubbles on the right side of the rating form to match your circled responses. For example, if you circled "a" and "d" for student no. 01, then you should darken the two bubbles marked "a" and "d" in row no. 1.

THANKS FOR YOUR COOPERATION!

NAME _____ SUBJECT _____

STUDENT	RATING ON OBJECTIVE: MASTER	RATING ON OBJECTIVE: NON-MASTER	CERTAINTY OF RATING: UN-CERTAIN	CERTAINTY OF RATING: FAIRLY CERTAIN	CERTAINTY OF RATING: VERY CERTAIN
1	a	b	c	d	e
2	a	b	c	d	e
3	a	b	c	d	e
4	a	b	c	d	e
5	a	b	c	d	e
6	a	b	c	d	e
7	a	b	c	d	e
8	a	b	c	d	e
9	a	b	c	d	e
10	a	b	c	d	e
11	a	b	c	d	e
12	a	b	c	d	e
13	a	b	c	d	e
14	a	b	c	d	e
15	a	b	c	d	e
16	a	b	c	d	e
17	a	b	c	d	e
18	a	b	c	d	e
19	a	b	c	d	e
20	a	b	c	d	e
21	a	b	c	d	e
22	a	b	c	d	e
23	a	b	c	d	e
24	a	b	c	d	e
25	a	b	c	d	e
26	a	b	c	d	e
27	a	b	c	d	e
28	a	b	c	d	e
29	a	b	c	d	e
30	a	b	c	d	e
31	a	b	c	d	e
32	a	b	c	d	e
33	a	b	c	d	e
34	a	b	c	d	e
35	a	b	c	d	e

USE NUMBER 2 PENCIL • ERASE COMPLETELY TO CHANGE

1	a	b	c	d	e
2	a	b	c	d	e
3	a	b	c	d	e
4	a	b	c	d	e
5	a	b	c	d	e
6	a	b	c	d	e
7	a	b	c	d	e
8	a	b	c	d	e
9	a	b	c	d	e
10	a	b	c	d	e
11	a	b	c	d	e
12	a	b	c	d	e
13	a	b	c	d	e
14	a	b	c	d	e
15	a	b	c	d	e
16	a	b	c	d	e
17	a	b	c	d	e
18	a	b	c	d	e
19	a	b	c	d	e
20	a	b	c	d	e
21	a	b	c	d	e
22	a	b	c	d	e
23	a	b	c	d	e
24	a	b	c	d	e
25	a	b	c	d	e
26	a	b	c	d	e
27	a	b	c	d	e
28	a	b	c	d	e
29	a	b	c	d	e
30	a	b	c	d	e
31	a	b	c	d	e
32	a	b	c	d	e
33	a	b	c	d	e
34	a	b	c	d	e
35	a	b	c	d	e
36	a	b	c	d	e
37	a	b	c	d	e
38	a	b	c	d	e
39	a	b	c	d	e
40	a	b	c	d	e
41	a	b	c	d	e
42	a	b	c	d	e
43	a	b	c	d	e
44	a	b	c	d	e
45	a	b	c	d	e
46	a	b	c	d	e
47	a	b	c	d	e
48	a	b	c	d	e
49	a	b	c	d	e
50	a	b	c	d	e

WRONG RESCORE

TABLE 3

CONSISTENCY STATISTICS FOR SETTING MASTERY STANDARDS

(UNWEIGHTED)

Test Identification: WS6.0B

Grades Tested: Grade 4
No. Cases 89

PROPORTION CORRECT FOR MASTERY	PROPORTION TRUE POSITIVE DECISIONS	PROPORTION TRUE NEGATIVE DECISIONS	PROPORTION FALSE POSITIVE DECISIONS	PROPORTION FALSE NEGATIVE DECISIONS	P INDEX OF DECISION CONSISTENCY
.10	.3034	.0449	.6517	.0000	.3483
.20	.3034	.0449	.6517	.0000	.3483
.30	.3034	.0674	.6292	.0000	.3708
.40	.3034	.0787	.6180	.0000	.3820
.50	.2921	.1348	.5618	.0112	.4270
.60	.2697	.2247	.4719	.0337	.4944
.70	.2022	.3933	.3034	.1011	.5955
.80	.1461	.5506	.1461	.1573	.6966
.90	.1011	.6629	.0337	.2022	.7640
1.00	.0225	.6966	.0000	.2809	.7191

TABLE 4

CONSISTENCY STATISTICS FOR SETTING MASTERY STANDARDS

(WEIGHTED FOR CERTAINTY OF MASTERY RATINGS)

Test Identification: WS6.0B

Grades Tested: Grade 4

No. Cases: 89

PROPORTION CORRECT FOR MASTERY	PROPORTION TRUE POSITIVE DECISIONS	PROPORTION TRUE NEGATIVE DECISIONS	PROPORTION FALSE POSITIVE DECISIONS	PROPORTION FALSE NEGATIVE DECISIONS	P INDEX OF DECISION CONSISTENCY
.10	.2624	.0446	.6931	.0000	.3069
.20	.2624	.0446	.6931	.0000	.3069
.30	.2624	.0743	.6634	.0000	.3366
.40	.2624	.0842	.6535	.0000	.3465
.50	.2475	.1485	.5891	.0149	.3960
.60	.2277	.2426	.4950	.0347	.4703
.70	.1832	.4257	.3119	.0792	.6089
.80	.1337	.5990	.1386	.1287	.7327
.90	.0891	.7030	.0347	.1733	.7921
1.00	.0149	.7376	.0000	.2475	.7525

CONSISTENCY OF MASTERY DECISIONS

FOR VARIABLE TEST LENGTH

(BASED ON BETA-BINOMIAL MODEL)

Test Identification:		WS6.0B
Grades Tested:		Grade 4
No. Cases:		89 (sample data)
No. Items on Pilot Test:		10
Mean on Pilot Test:		6.3146
Standard Deviation on Pilot Test:		2.1511
Beta-Binomial Parameters:	Alpha:	5.1189
	Beta:	2.9875
No. of Items in Test Bank:		20
Proportion Correct for Mastery:		.90

NUMBER ITEMS	CUTOFF SCORE	MEAN	STANDARD DEVIATION	P INDEX
2	2	1.2629	.7187	.5884
3	3	1.8944	.9228	.6564
4	4	2.5258	1.1125	.7237
5	5	3.1573	1.2941	.7789
6	6	3.7888	1.4707	.8220
7	7	4.4202	1.6439	.8553
8	8	5.0517	1.8146	.8812
9	9	5.6831	1.9836	.9014
10	9	6.3146	2.1511	.8064
11	10	6.9461	2.3176	.8305
12	11	7.5775	2.4831	.8511
13	12	8.2090	2.6480	.8689
14	13	8.8404	2.8123	.8841
15	14	9.4719	2.9762	.8971
16	15	10.1034	3.1396	.9084
17	16	10.7348	3.3026	.9181
18	17	11.3663	3.4654	.9266
19	18	11.9977	3.6279	.9339
20	18	12.6292	3.7902	.8924

TABLE 6

CONSISTENCY OF MASTERY DECISIONS
FOR VARIABLE TEST LENGTH

(BASED ON BETA-BINOMIAL MODEL)

Test Identification:		WS6.0B
Grades Tested:		Grade 4
No. Cases:		1000 (simulated data)
No. Items on Pilot Test:		10
Mean on Pilot Test:		5.0379
Standard Deviation on Pilot Test:		1.6820
Beta-Binomial Parameters:	Alpha:	33.9200
	Beta:	33.4096
No. of Items in Test Bank:		20
Proportion Correct for Mastery:		.90

NUMBER ITEMS	CUTOFF SCORE	MEAN	STANDARD DEVIATION	P INDEX
2	2	1.0076	.7122	.6251
3	3	1.5114	.8786	.7733
4	4	2.0152	1.0217	.8720
5	5	2.5190	1.1503	.9293
6	6	3.0227	1.2687	.9612
7	7	3.5265	1.3797	.9786
8	8	4.0303	1.4848	.9881
9	9	4.5341	1.5853	.9934
10	9	5.0379	1.6820	.9681
11	10	5.5417	1.7755	.9805
12	11	6.0455	1.8662	.9881
13	12	6.5493	1.9546	.9927
14	13	7.0531	2.0410	.9955
15	14	7.5569	2.1256	.9972
16	15	8.0606	2.2086	.9983
17	16	8.5644	2.2902	.9989
18	17	9.0682	2.3705	.9993
19	18	9.5720	2.4497	.9996
20	18	10.0758	2.5278	.9982

REFERENCES

Anderson, L. W. Unpublished lecture notes on "Educational Measurement". College of Education, University of South Carolina, 1977.

Byars, A. G. Performance indices of the "Individualized Criterion-Referenced Tests". Paper presented at the third annual convention of the Eastern Educational Research Association, 1979.

Garcia-Quintana, R. A. Procedures for setting multiple cutoff scores in the South Carolina Basic Skills Assessment Program based on discriminant analysis and a consideration of the P index. Manuscript in preparation, Office of Research, South Carolina Department of Education, 1979.

Huynh, H. Statistical Consideration of mastery scores. *Psychometrika*, 1976a, *41*, 65-78.

Huynh, H. On the reliability of decisions in domain-referenced testing. *Journal of Educational Measurement*, 1976b, *13*, 253-264.

Huynh, H. Two simple classes of mastery scores based on the beta-binomial model. *Psychometrika*, 1977, *42*, 601-608.

Huynh, H. Unpublished lecture notes on "Technical Issues in Testing and Measurement". College of Education, University of South Carolina, 1978a.

Huynh, H. Computation and inference for two reliability indices in mastery testing based on the beta-binomial model. Research Memorandum 78-1, Paper presented at the 17th Annual Southeastern Invitational Conference on Measurement in Education, 1978b.

Huynh, H. Personal communication. College of Education, University of South Carolina, 1979.

Huynh, H. and Saunders, J. C. Bayesian and empirical Baayes approaches to setting mastery scores. Research Memorandum 79-2, Educational Research Program, College of Education, University of South Carolina, 1979.

Siegel, S. *Nonparametric Statistics*. New York: McGraw-Hill Book Co., 1956.

Subkoviak, M. J. Empirical investigation of procedures for estimating reliability for mastery tests. *Journal of Educational Measurement*, 1978, *15*, 111-116.

Thorndike, R. L. *Educational Measurement*. Washington, D.C.: American Council on Education, 1971.

Wright, B. D. Solving measurement problems with the Rasch model. *Journal of Educational Measurement*, 1977, *14*, 97-116.